

# Assessing cumulative evidence: Venice criteria update

HuGENet Workshop January 2008

John P.A. Ioannidis

*Professor and Chairman, Department of Hygiene and Epidemiology*

*University of Ioannina School of Medicine, Ioannina, Greece*

*Professor of Medicine (adjunct), Tufts University School of Medicine, Boston, USA*

- Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD, Higgins JP, McCarthy MI, McDermott DH, Page GP, Rebbeck TR, Seminara D, Khoury MJ. Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol*. 2008 Feb;37(1):120-32. Epub 2007 Sep 26.

# Grading the evidence: the Venice criteria (IJE, 2007)



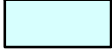
AAA	ABA	ACA
AAB	ABB	ACB
AAC	ABC	ACC

First letter = amount

Second letter = replication

Third letter = protection from bias

BAA	BBA	BCA
BAB	BBB	BCB
BAC	BBC	BCC

	Strong evidence
	Moderate evidence
	Weak evidence

CAA	CBA	CCA
CAB	CBB	CCB
CAC	CBC	CCC

# The three criteria

**Table 1** Considerations for epidemiologic credibility in the assessment of cumulative evidence on genetic associations

Criteria	Categories	Proposed operationalization
Amount of evidence	<p>A: Large-scale evidence</p> <p>B: Moderate amount of evidence</p> <p>C: Little evidence</p>	<p>Thresholds may be defined based on sample size, power or false-discovery rate considerations. The frequency of the genetic variant of interest should be accounted for. As a simple rule, we suggest that category A requires over 1000 subjects (total number of cases and controls assuming 1:1 ratio) evaluated in the least common genetic group of interest; B corresponds to 100–1000 subjects evaluated in this group and C corresponds to &lt;100 subjects evaluated in this group (see ‘Discussion’ section in the text and Table 2 for further elaboration).<sup>a</sup></p>
Replication	<p>A: Extensive replication including at least one well-conducted meta-analysis with little between-study inconsistency</p> <p>B: Well-conducted meta-analysis with some methodological limitations or moderate between-study inconsistency</p> <p>C: No association; no independent replication; failed replication; scattered studies; flawed meta-analysis or large inconsistency</p>	<p>Between-study inconsistency entails statistical considerations (e.g. defined by metrics such as <math>I^2</math>, where values of 50% and above are considered large and values of 25–50% are considered moderate inconsistency) and also epidemiological considerations for the similarity/standardization or at least harmonization of phenotyping, genotyping and analytical models across studies. See ‘Discussion’ section in the text for the threshold (statistical or others) required for claiming replication under different circumstances (e.g. with or without including the discovery data in situations with massive testing of polymorphisms).</p>
Protection from bias	<p>A: Bias, if at all present, could affect the magnitude but probably not the presence of the association</p> <p>B: No obvious bias that may affect the presence of the association but there is considerable missing information on the generation of evidence</p> <p>C: Considerable potential for or demonstrable bias that can affect even the presence or absence of the association</p>	<p>A prerequisite for A is that the bias due to phenotype measurement, genotype measurement, confounding (population stratification) and selective reporting (for meta-analyses) can be appraised as not being high (as shown in detail in Table 3) plus there is no other demonstrable bias in any other aspect of the design, analysis or accumulation of the evidence that could invalidate the presence of the proposed association. In category B, although no strong biases are visible, there is no such assurance that major sources of bias have been minimized or accounted for because information is missing on how phenotyping, genotyping and confounding have been handled. Given that occult bias can never be ruled out completely, note that even in category A, we use the qualifier ‘probably’.</p>

<sup>a</sup>For example, if the association pertains to the presence of homozygosity for a common variant and if the frequency of homozygosity is 3%, then category A amount of evidence requires over 30 000 subjects and category B between 3000 and 30 000.

# Amount of evidence

- A: Large-scale evidence
- B: Moderate amount of evidence
- C: Little evidence

Thresholds may be defined based on sample size, power or false-discovery rate considerations. The frequency of the genetic variant of interest should be accounted for. As a simple rule, we suggest that category A requires over 1000 subjects (total number of cases and controls assuming 1:1 ratio) evaluated in the least common genetic group of interest; B corresponds to 100–1000 subjects evaluated in this group and C corresponds to <100 subjects evaluated in this group (see 'Discussion' section in the text and Table 2 for further elaboration).<sup>a</sup>

# Options of amount of evidence

- Simple operational: sample size of the least common genetic group among those compared (it could reflect participants or alleles, depending on the model)
- Power
- False-discovery rate

**Table 2** Power calculations for associations with  $n_{\text{minor}} = 1000$  for various ORs and various frequencies of the minor genetic group ( $f_{\text{minor}}$ )<sup>a</sup>

OR	$f_{\text{minor}}$	Power for $\alpha = 0.05$	Power for $\alpha = 10^{-7}$
1.10	0.01	0.32	<0.001
1.20	0.01	0.82	0.007
1.30	0.01	0.98	0.12
1.50	0.01	1.00	0.83
2.00	0.01	1.00	1.00
5.00	0.01	1.00	1.00
1.10	0.05	0.31	<0.001
1.20	0.05	0.80	0.006
1.30	0.05	0.98	0.09
1.50	0.05	1.00	0.78
2.00	0.05	1.00	1.00
5.00	0.05	1.00	1.00
1.10	0.10	0.30	<0.001
1.20	0.10	0.78	0.005
1.30	0.10	0.97	0.74
1.50	0.10	1.00	1.00
2.00	0.10	1.00	1.00
5.00	0.10	1.00	1.00
1.10	0.25	0.25	<0.001
1.20	0.25	0.69	0.002
1.30	0.25	0.94	0.04
1.50	0.25	1.00	0.52
2.00	0.25	1.00	1.00
5.00	0.25	1.00	1.00
1.10	0.50	0.18	<0.001
1.20	0.50	0.51	<0.001
1.30	0.50	0.81	0.006
1.50	0.50	0.99	0.15
2.00	0.50	1.00	0.96
5.00	0.50	1.00	1.00

<sup>a</sup>All calculations assume the same number of cases and controls; results are relatively robust to modest deviations in the allocation ratio. The minor genetic group is the smallest of the two groups contrasted and may have been selected based on genotype or allele considerations.

# How does it look so far

- For candidate-gene era, several postulated variants fail to reach “Amount A”
- With large collaborative efforts currently, this is typically not a problem for common variants with frequency  $>5-10\%$
- For variants with lesser frequency, very large sample sizes may be required, e.g. for  $f=1\%$ , we need  $n=100,000$



# Replication

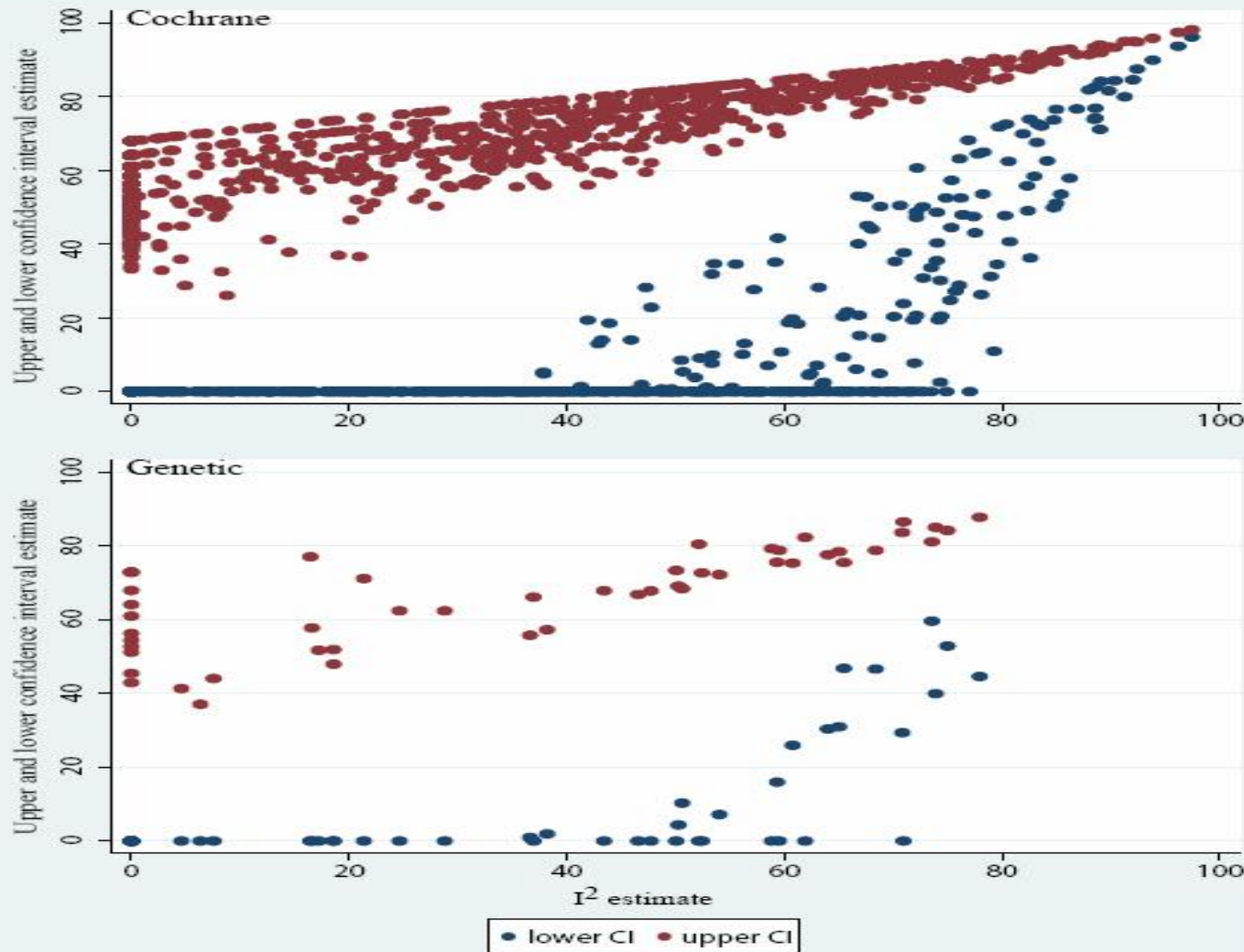
- A: Extensive replication including at least one well-conducted meta-analysis with little between-study inconsistency
- B: Well-conducted meta-analysis with some methodological limitations or moderate between-study inconsistency
- C: No association; no independent replication; failed replication; scattered studies; flawed meta-analysis or large inconsistency

Between-study inconsistency entails statistical considerations (e.g. defined by metrics such as  $I^2$ , where values of 50% and above are considered large and values of 25–50% are considered moderate inconsistency) and also epidemiological considerations for the similarity/standardization or at least harmonization of phenotyping, genotyping and analytical models across studies. See 'Discussion' section in the text for the threshold (statistical or others) required for claiming replication under different circumstances (e.g. with or without including the discovery data in situations with massive testing of polymorphisms).

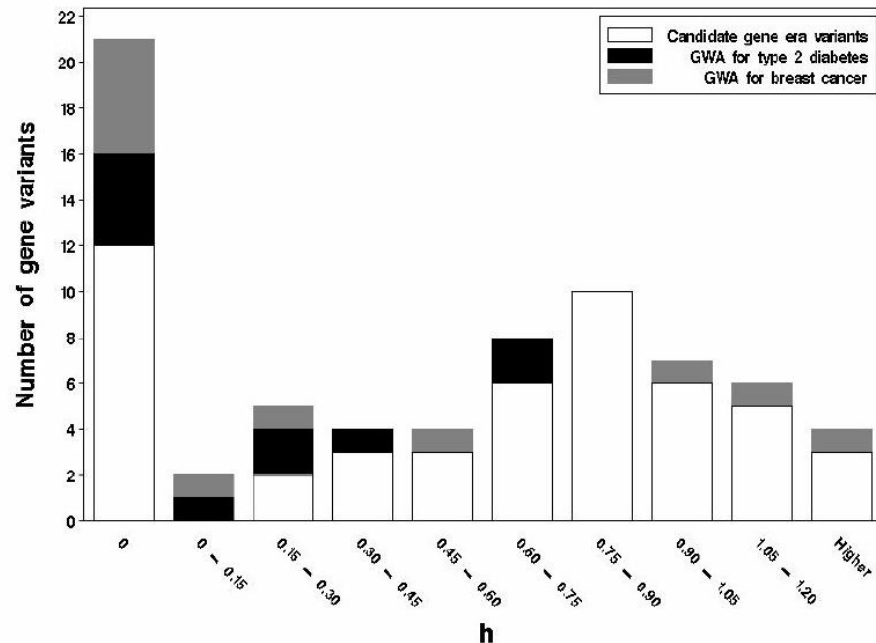
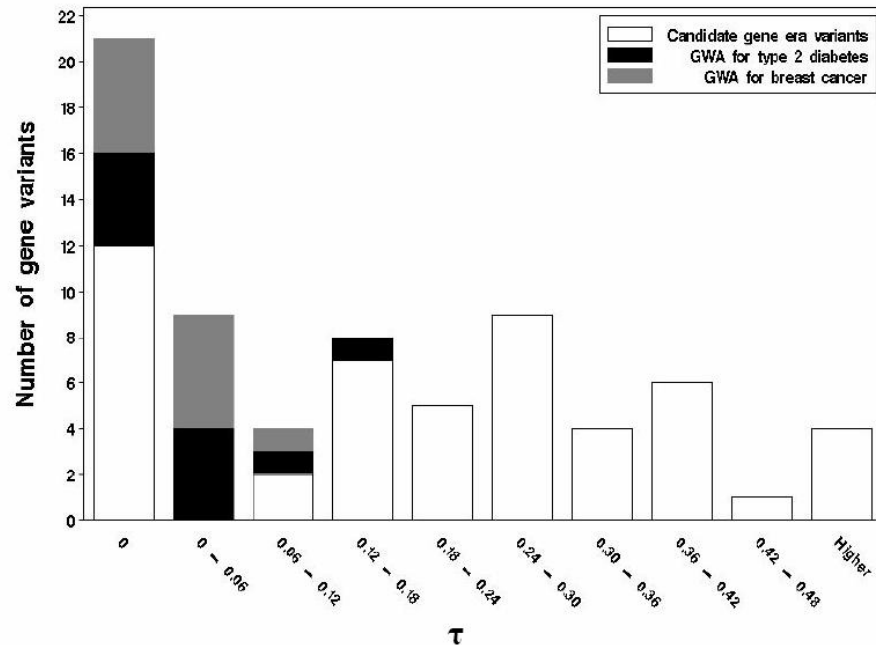
# Any problems with C?

- No association: operational definition of no association so far based on traditional  $p > 0.05$  for total replication efforts; beware that minute effects can never be excluded. The Venice criteria should not be used to tell that we have strong evidence for an association to be *excluded*.
- No independent replication: obvious
- Failed replication: same considerations as “no association”
- Scattered studies: in the absence of better type/collection of evidence
- Flawed meta-analysis: judgment call; have not been invoked to-date to kill any association, as far as I know
- Large inconsistency: to discuss further, since “B” also mentions “moderate between-study inconsistency”

# Uncertainty of $I^2$ estimates of heterogeneity in meta-analyses



# Heterogeneity in candidate gene era and GWA era





# Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations

John P. A. Ioannidis<sup>1,2,3\*</sup>, Nikolaos A. Patsopoulos<sup>1</sup>, Evangelos Evangelou<sup>1</sup>

**1** Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, **2** Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece, **3** Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts, United States of America

**Background.** Meta-analysis is the systematic and quantitative synthesis of effect sizes and the exploration of their diversity across different studies. Meta-analyses are increasingly applied to synthesize data from genome-wide association (GWA) studies and from other teams that try to replicate the genetic variants that emerge from such investigations. Between-study heterogeneity is important to document and may point to interesting leads. **Methodology/Principal Findings.** To exemplify these issues, we used data from three GWA studies on type 2 diabetes and their replication efforts where meta-analyses of all data using fixed effects methods (not incorporating between-study heterogeneity) have already been published. We considered 11 polymorphisms that at least one of the three teams has suggested as susceptibility loci for type 2 diabetes. The  $I^2$  inconsistency metric (measuring the amount of heterogeneity not due to chance) was different from 0 (no detectable heterogeneity) for 6 of the 11 genetic variants; inconsistency was moderate to very large ( $I^2 = 32\text{--}77\%$ ) for 5 of them. For these 5 polymorphisms, random effects calculations incorporating between-study heterogeneity revealed more conservative p-values for the summary effects compared with the fixed effects calculations. These 5 associations were perused in detail to highlight potential explanations for between-study heterogeneity. These include identification of a marker for a correlated phenotype (e.g. *FTO* rs8050136 being associated with type 2 diabetes through its effect on obesity); differential linkage disequilibrium across studies of the identified genetic markers with the respective culprit polymorphisms (e.g., possibly the case for *CDKAL1* polymorphisms or for rs9300039 and markers in linkage disequilibrium, as shown by additional studies); and potential bias. Results were largely similar, when we treated the discovery and replication data from each GWA investigation as separate studies. **Significance.** Between-study heterogeneity is useful to document in the synthesis of data from GWA investigations and can offer valuable insights for further clarification of gene-disease associations.

Citation: Ioannidis JPA, Patsopoulos NA, Evangelou E (2007) Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. PLoS ONE 2(9): e841. doi:10.1371/journal.pone.0000841

**Table 1.** Between-study heterogeneity and random versus fixed effects calculations for polymorphisms that were considered "confirmed"

GENE	Polymorphism	Q (p)	I <sup>2</sup> (95% CI)	Random effects OR (95% CI)	Fixed effects OR (95% CI)	Random effects p-value	Fixed effects p-value
—	rs9300039 <sup>a</sup>	7.98 (0.019)	75% (0–90)	1.25 (1.04–1.50)	1.25 (1.15–1.37)	0.015	4.3 × 10 <sup>-7</sup>
<i>FTO</i>	rs8050136	8.62 (0.013)	77% (0–91)	1.13 (1.02–1.25)	1.17 (1.12–1.22)	0.015	1.3 × 10 <sup>-12</sup>
<i>PPARG</i>	rs1801282	3.80 (0.15)	47% (0–84)	1.16 (1.07–1.25)	1.14 (1.08–1.20)	0.0003	1.7 × 10 <sup>-6</sup>
<i>CDKAL1</i>	rs10946398 <sup>b</sup>	3.73 (0.16)	46% (0–84)	1.12 (1.07–1.17)	1.12 (1.08–1.16)	3.2 × 10 <sup>-6</sup>	4.1 × 10 <sup>-11</sup>
<i>SLC30A8</i>	rs13266634	2.92 (0.23)	32% (0–81)	1.12 (1.07–1.18)	1.12 (1.07–1.16)	8.7 × 10 <sup>-6</sup>	5.3 × 10 <sup>-8</sup>
<i>CDKN2B</i>	rs564398	1.48 (0.48)	0% (0–73)	1.12 (1.07–1.17)	1.12 (1.07–1.17)	1.2 × 10 <sup>-7</sup>	1.2 × 10 <sup>-7</sup>
<i>HHEX</i>	rs5015480– rs1111875	0.45 (0.80)	0% (0–73)	1.13 (1.08–1.17)	1.13 (1.08–1.17)	5.7 × 10 <sup>-10</sup>	5.7 × 10 <sup>-10</sup>
<i>KCNJ11</i>	rs5215 <sup>c</sup>	0.56 (0.76)	0% (0–73)	1.14 (1.10–1.19)	1.14 (1.10–1.19)	5 × 10 <sup>-11</sup>	5 × 10 <sup>-11</sup>
<i>IGF2BP2</i>	rs4402960	2.65 (0.27)	25% (0–79)	1.15 (1.10–1.19)	1.14 (1.10–1.18)	6.5 × 10 <sup>-12</sup>	8.6 × 10 <sup>-16</sup>
<i>CDKN2B</i>	rs10811661	0.03 (0.99)	0% (0–73)	1.20 (1.14–1.25)	1.20 (1.14–1.25)	7.8 × 10 <sup>-15</sup>	7.8 × 10 <sup>-15</sup>
<i>TCF7L2</i>	rs7901695 <sup>d</sup>	0.24 (0.89)	0% (0–73)	1.37 (1.31–1.43)	1.37 (1.31–1.43)	1.0 × 10 <sup>-48</sup>	1.0 × 10 <sup>-48</sup>

Additive models are presented, as in the main analyses of the original papers. Fixed effects calculations are Mantel-Haenszel estimates as in the original papers. Random effects calculations use the DerSimonian and Laird estimators for the between-study variance.

CI: confidence interval; OR: odds ratio

<sup>a</sup>multi-marker tag in DGI and rs1514823 in the UK study

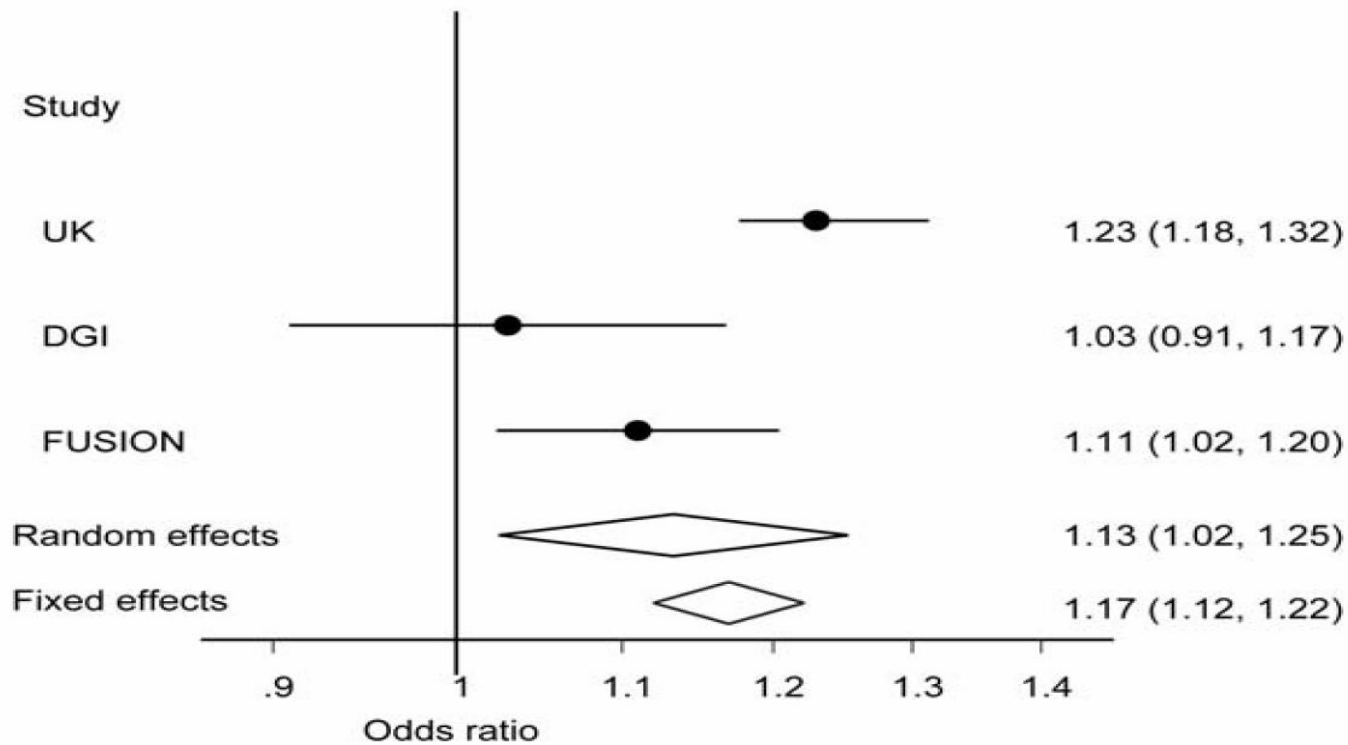
<sup>b</sup>rs7754840 in FUSION

<sup>c</sup>rs5219 in FUSION and DGI

<sup>d</sup>rs7903146 in FUSION and DGI

doi:10.1371/journal.pone.0000841.t001

# An inconsistent association mirroring a different association: *FTO*, type 2 diabetes, and obesity



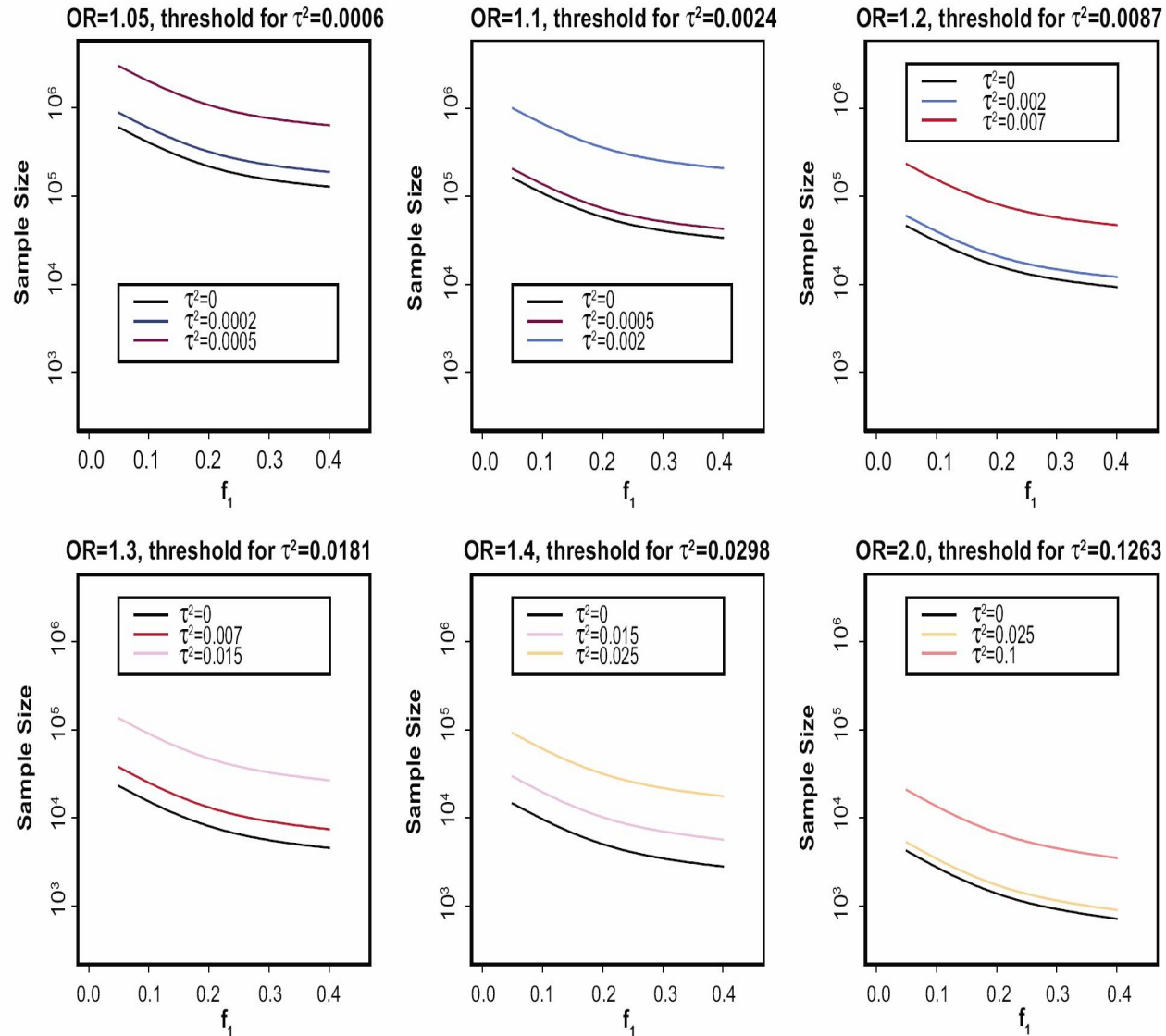
# Impact of criterion 2 on FTO variant

- The variant has weak epidemiological support for an association with type 2 diabetes
- It has strong epidemiological support for an association with obesity



# Inconsistency and non-replicability threshold

- Inconsistency may be due to either bias or genuine between-study heterogeneity
- Beyond a given threshold of inconsistency, no matter how large studies we conduct, we may never have enough power to replicate an association (non-replicability threshold)



**Fig. 2.** Mean sample sizes required to detect odds ratios of 1.05, 1.1, 1.2, 1.3, 1.4, and 2.0 with power 80% at  $\alpha = 0.0000001$  as a function of genotype frequency  $f_1$  for a metaanalysis of 10 equally large studies.

# Protection from bias

- A: Bias, if at all present, could affect the magnitude but probably not the presence of the association
- B: No obvious bias that may affect the presence of the association but there is considerable missing information on the generation of evidence
- C: Considerable potential for or demonstrable bias that can affect even the presence or absence of the association

A prerequisite for A is that the bias due to phenotype measurement, genotype measurement, confounding (population stratification) and selective reporting (for meta-analyses) can be appraised as not being high (as shown in detail in Table 3) plus there is no other demonstrable bias in any other aspect of the design, analysis or accumulation of the evidence that could invalidate the presence of the proposed association. In category B, although no strong biases are visible, there is no such assurance that major sources of bias have been minimized or accounted for because information is missing on how phenotyping, genotyping and confounding have been handled. Given that occult bias can never be ruled out completely, note that even in category A, we use the qualifier 'probably'.

**Table 3** Typical biases and their typical impact on associations depending on the status of the evidence

Biases	Status of the evidence	Likelihood of bias to invalidate an observed association		
		Small OR <1.15	Typical OR 1.15–1.8	Large OR >1.8
Bias in phenotype definition	Not reported what was done	Unknown	Unknown	Unknown
	Unclear phenotype definitions	Possible/High	Possible/High	Possible/High
	Clear widely agreed definitions of phenotypes	Low/None	Low/None	Low/None
	Efforts for retrospective harmonization	Possible/High	Low	Low/None
	Prospective standardization of phenotypes	Low/None	Low/None	Low/None
Bias in genotyping	Not reported what was done	Unknown	Unknown	Unknown
	No quality control checks	Possible/High	Low	Low
	Appropriate quality control checks	Low	Low	Low/None
Population stratification	Not reported what was done	Unknown	Unknown	Unknown
	Nothing done <sup>a</sup>	Possible/High	Possible/High	Possible/High
	Same descent group <sup>b</sup>	Possible/High	Low	Low/None
	Adjustment for reported descent	Possible/High	Low	Low/None
	Family-based design	Low/None	Low/None	Low/None
	Genomic control, PCA or similar method	Low/None	Low/None	Low/None
Selective reporting biases	Meta-analysis of published data	Possible/High	Possible	Possible
	Retrospective efforts to include unpublished data	Possible/High	Possible	Possible
	Meta-analysis within consortium	Low/None	Low/None	Low/None

A research finding cannot reach  
credibility over 50% unless

$$u < R$$

i.e. bias must be less than the pre-study  
odds

# Bias checks for retrospective meta-analysis

## “Automated checks”

- Effect size  $<1.15$ -fold from the null effect
- Association lost with exclusion of first study
- Association lost with exclusion of HWE-violating studies or with adjustment for HWE
- Evidence for small-study effect in an asymmetry regression test with proper type I error (e.g. Harbord, Stat Med)
- Evidence for excess of single studies with formally statistically significant results (Ioannidis and Trikalinos, Clinical Trials)

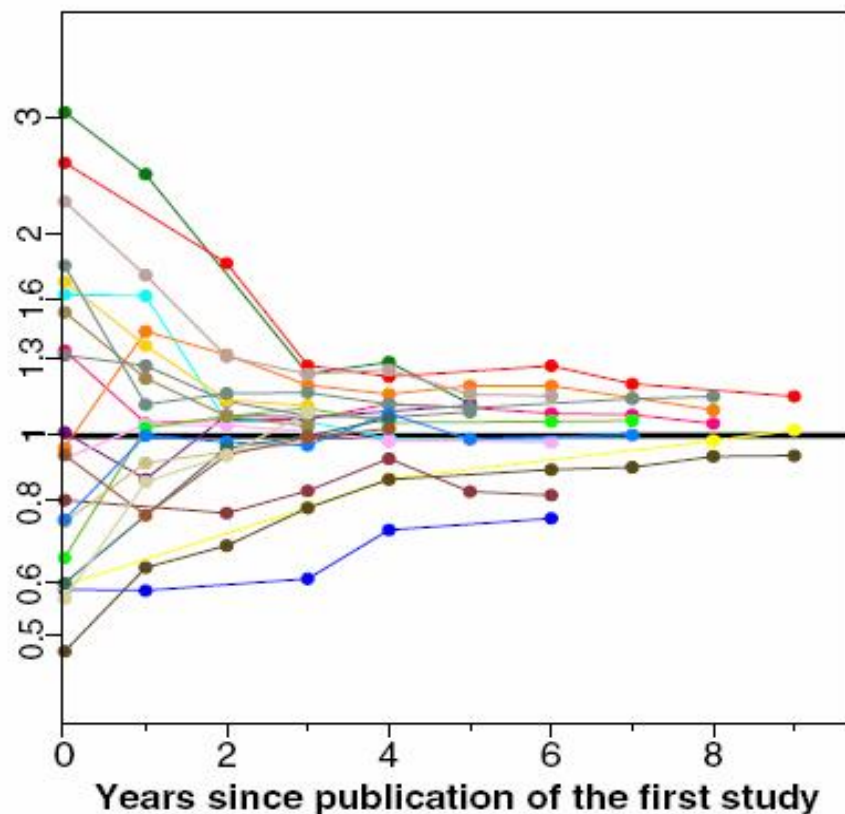
## “Consider whether they are problems”

- Unclear/misclassified phenotypes with possible differential misclassification against genotyping
- Differential misclassification of genotyping against phenotypes
- Major concerns for population stratification (need to justify for affecting  $OR > 1.15$ -fold, not invoked to-date)
- Any other reason (case-by-case basis) that would destroy the association

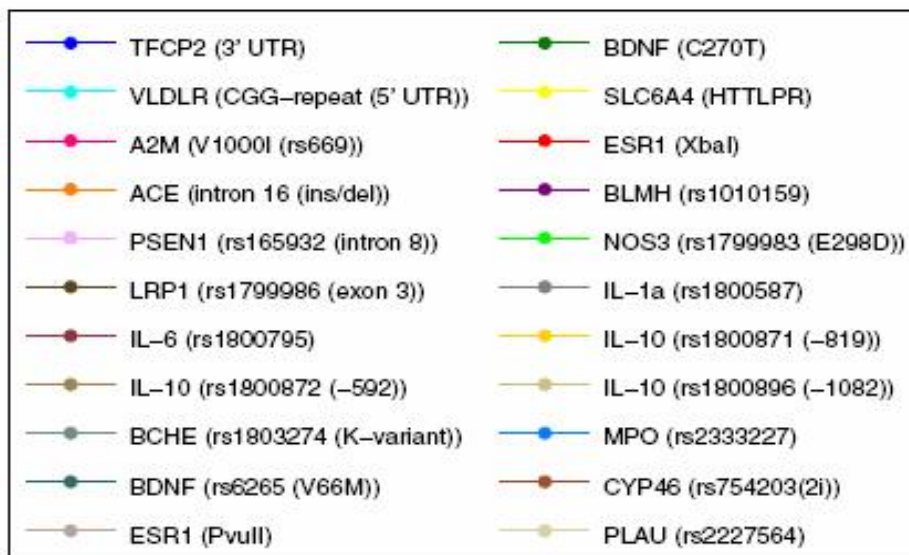
# Bias checks for a prospective consortium analysis

- Magnitude of effect size, small-study effects, excess of studies with significant findings are not an issue here, provided there is no selective reporting (basic trust)
- The other considerations still need to be raised





Meta-analyses with a significant excess of significant single studies in Alzheimer's disease genetics: genuine heterogeneity or bias?





# Still to debate

- Threshold for association/no association: so far we are using  $p < 0.05$  (that survives after excluding the discovery component in the data)
- Obviously, this is very lenient
- Beware though that associations that manage to get grade A for amount of evidence, and don't get much lower p-values than this, have either very small effects (and get a C for protection from bias if retrospective meta-analysis) or moderate/large heterogeneity (and get a B or C for replication consistency)
- Rarely a  $p > 10^{-5}$  excluding discovery data gets “strong epidemiological evidence” grading
- Instead of trying to play with the threshold of p-value for claiming an association or not, one may keep the lenient 0.05 and for those variants that do succeed against this lenient threshold and also get an overall AAA (strong epidemiological evidence) grading, try to add a credibility estimate based on Bayesian considerations

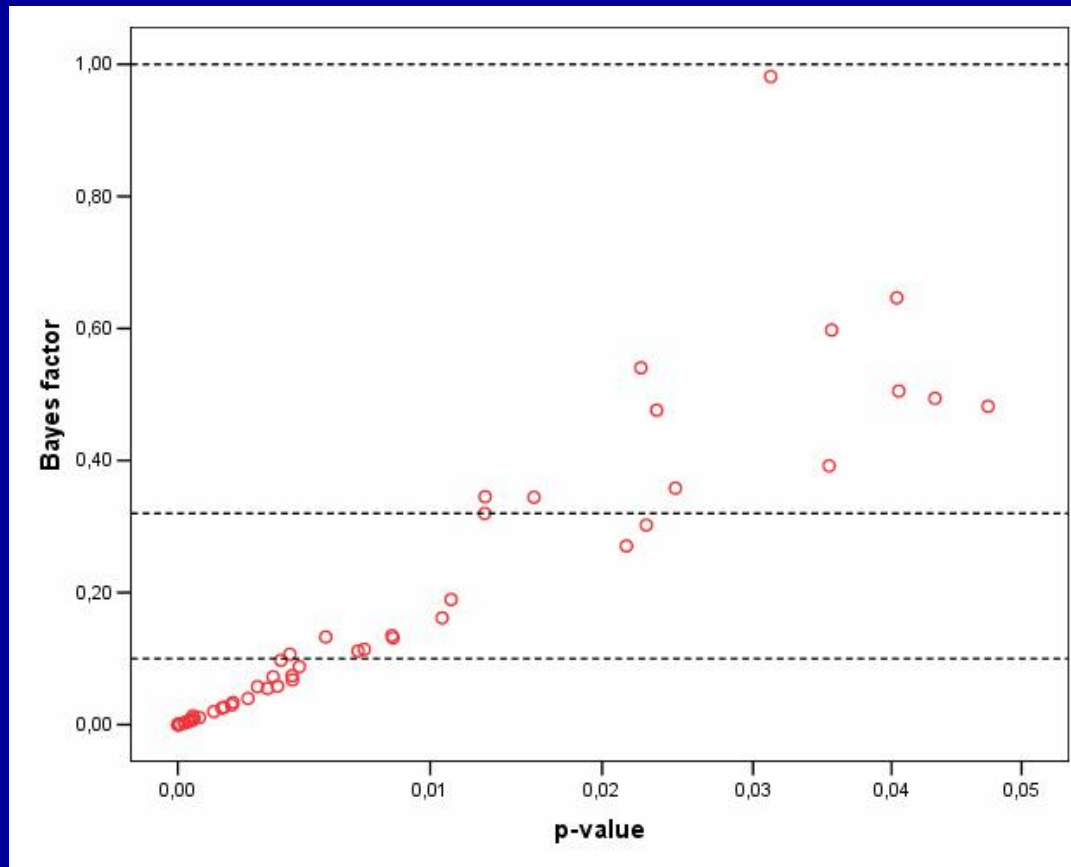
# Calibration of credibility

$$B = \sqrt{(1 + (m / n_0))} \exp[(-z_m^2) / (2(1 + (n_0 / m)))]$$

$$n_0 = 2\sigma^2 / (\pi\theta_A^2) = 2m \text{var}(\theta) / (\pi\theta_A^2)$$

$$n_0 / m = 2 \text{var}(\theta) / (\pi\theta_A^2)$$

# Calibration of credibility: genetic meta-analyses



# Evolving credibility in successive genetic meta-analyses

Earlier M-A (author and year)	Gene (variant); Contrast	Disease	OR (95% CI) in M-A	OR (95% CI) M-A2	M-A2 (author and year)	Differences	Bayes
No substantial support							
Boekholdt 2001	<i>FGB</i> / <i>FGB</i> promoter (455G/A); AA vs GG	MI	1.46 (1.00, 2.13)	1.12 (0.90, 1.41)	Smith 2005	Allele/wider	0.48/NP
Maraganore 2004	<i>UCH-L1</i> (S18Y); S/S vs. other	Parkinson	1.20 (1.02, 1.40)	0.96 (0.86, 1.08)	Healy 2006	None/None	0.48/NP
Kosmas 2004	<i>MTHFR</i> (677C/T); TT vs. other	Preeclampsia	1.21 (1.01, 1.45)	1.01 (0.79, 1.29)	Lin 2005	None/None	0.60/NP
Burzotta 2004	<i>F2</i> (20210G/A); other vs. GG	MI	1.32 (1.01, 1.72)	1.25 (1.05, 1.50)	Ye 2006	Allele	0.51/0.28
Jonsson 2003	<i>DRD3</i> (Ser9Gly) SerSer vs. other	Schizophrenia	1.10 (1.01, 1.21)	1.05 (0.97, 1.13)	Jonsson 2004	None/None	0.98/NP
Combarros 2003	<i>IL1A</i> (-889); 2/2 vs. Other	Alzheimer	2.35 (1.03, 5.37)	1.08 (0.98, 1.18)	Bertram 2007	Allele/wider	0.49/NP

**Table 1.** Estimated Bayes factors for selected associations proposed by GWA studies according to different values of  $\theta_A$  (0.049, 0.140, 0.262, 0.405, and 0.588, corresponding to odds ratios of 1.05, 1.15, 1.30, 1.50 and 1.80, respectively)

GENE	Variant	OR (95% CI)	p-value	Estimated $-\log_{10}$ (Bayes factor) under different assumptions for the $\theta_A$				
				$\theta_A=0.049$	$\theta_A=0.140$	$\theta_A=0.262$	$\theta_A=0.405$	$\theta_A=0.588$
<b>Periodic limb movements in sleep</b>								
<i>BTBD9</i>	rs3923809	1.72 (1.50-1.98)	3x10 <sup>-14</sup>	5.26	10.35	11.30	11.44	11.40
<b>Type 2 diabetes mellitus</b>								
---	rs9300039	1.25 (1.04-1.50)	0.015	0.31	0.67	0.63	0.50	0.36
<i>FTO</i>	rs8050136	1.13 (1.02-1.25)	0.015	0.56	0.63	0.45	0.28	0.12
<i>PPARG</i>	rs1801282	1.16 (1.07-1.25)	0.0003	1.74	2.04	1.88	1.71	1.56
<i>CDKAL1</i>	rs10946398	1.12 (1.07-1.17)	3.2x10 <sup>-6</sup>	3.68	3.74	3.53	3.35	3.20
<i>SLC30A8</i>	rs13266634	1.12 (1.07-1.18)	8.7x10 <sup>-6</sup>	3.26	3.36	3.15	2.98	2.82
<i>CDKN2B</i>	rs564398	1.12 (1.07-1.17)	1.2x10 <sup>-7</sup>	4.89	5.09	4.90	4.72	4.57
<i>HHEX</i>	rs5015480- rs1111875	1.13 (1.08-1.17)	5.7x10 <sup>-10</sup>	7.01	7.29	7.10	6.93	6.78
<i>KCNJ11</i>	rs5215	1.14 (1.10-1.19)	5x10 <sup>-11</sup>	7.96	8.31	8.13	7.96	7.80
<i>IGF2BP2</i>	rs4402960	1.15 (1.10-1.19)	6.5x10 <sup>-12</sup>	8.75	9.17	8.99	8.82	8.67
<i>CDKN2B</i>	rs10811661	1.20 (1.14-1.25)	7.8x10 <sup>-15</sup>	10.99	12.00	11.90	11.75	11.60
<i>TCF7L2</i>	rs7901695	1.37 (1.31-1.43)	1.0x10 <sup>-48</sup>	>30	>30	>30	>30	>30
<b>Parkinson's disease</b>								
<i>SEMA5A</i>	rs7702187	1.74 (1.36–2.24)	7.62×10 <sup>-6</sup>	0.78	2.62	3.34	3.48	3.45
---	rs10200894	1.84 (1.38–2.45)	1.70×10 <sup>-5</sup>	0.57	2.17	2.96	3.15	3.15
---	rs2313982	2.01 (1.44–2.79)	1.79×10 <sup>-5</sup>	0.44	1.92	2.82	3.10	3.15
---	rs17329669	1.71 (1.33–2.21)	2.30×10 <sup>-5</sup>	0.67	2.29	2.93	3.05	3.01
---	rs7723605	1.78 (1.35–2.35)	3.30×10 <sup>-5</sup>	0.56	2.07	2.75	2.90	2.89
---	ss46548856	1.88 (1.38–2.57)	3.65×10 <sup>-5</sup>	0.45	1.86	2.64	2.85	2.86
<i>GALNT3</i>	rs16851009	1.84 (1.36–2.49)	4.17×10 <sup>-5</sup>	0.47	1.88	2.62	2.80	2.80
<i>PRDM2</i>	rs2245218	1.67 (1.29–2.14)	4.61×10 <sup>-5</sup>	0.62	2.11	2.69	2.78	2.73
<i>PASDI</i>	rs7878232	1.38 (1.17–1.62)	6.87×10 <sup>-5</sup>	1.12	2.44	2.62	2.56	2.45
---	rs1509269	1.71 (1.30–2.26)	9.21×10 <sup>-5</sup>	0.49	1.81	2.40	2.51	2.48
---	rs11737074	1.50 (1.21–1.86)	1.55×10 <sup>-4</sup>	0.68	1.96	2.30	2.29	2.21

**Table 2.** Credibility estimates for the associations of Table 1

<i>Gene</i>	Variant	With prior credibility $C_0=0.0001$			With prior credibility $C_0=0.00001$			With prior credibility $C_0=0.000001$		
		$\theta_A=0.049$	$\theta_A=0.262$	$\theta_A=0.588$	$\theta_A=0.049$	$\theta_A=0.262$	$\theta_A=0.588$	$\theta_A=0.049$	$\theta_A=0.262$	$\theta_A=0.588$
<b>Periodic limb movements in sleep</b>										
<i>BTBD9</i>	rs3923809	0.948	1.000	1.000	0.645	1.000	1.000	0.154	1.000	1.000
<b>Type 2 diabetes mellitus</b>										
---	rs9300039	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<i>FTO</i>	rs8050136	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<i>PPARG</i>	rs1801282	0.005	0.007	0.004	0.001	0.001	0.000	0.000	0.000	0.000
<i>CDKAL1</i>	rs10946398	0.325	0.252	0.136	0.046	0.033	0.015	0.005	0.003	0.002
<i>SLC30A8</i>	rs13266634	0.154	0.124	0.062	0.018	0.014	0.007	0.002	0.001	0.001
<i>CDKN2B</i>	rs564398	0.886	0.887	0.788	0.437	0.440	0.270	0.072	0.073	0.036
<i>HHEX</i>	rs5015480- rs1111875	0.999	0.999	0.998	0.990	0.992	0.984	0.911	0.927	0.857
<i>KCNJ11</i>	rs5215	1.000	1.000	1.000	0.999	0.999	0.998	0.989	0.993	0.985
<i>IGF2BP2</i>	rs4402960	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.999	0.998
<i>CDKN2B</i>	rs10811661	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>TCF7L2</i>	rs7901695	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<b>Parkinson's disease</b>										
<i>SEMA5A</i>	rs7702187	0.001	0.179	0.222	0.000	0.021	0.028	0.000	0.002	0.003
---	rs10200894	0.000	0.084	0.125	0.000	0.009	0.014	0.000	0.001	0.001
---	rs2313982	0.000	0.062	0.123	0.000	0.007	0.014	0.000	0.001	0.001
---	rs17329669	0.000	0.079	0.094	0.000	0.008	0.010	0.000	0.001	0.001
---	rs7723605	0.000	0.054	0.071	0.000	0.006	0.008	0.000	0.001	0.001
---	ss46548856	0.000	0.042	0.068	0.000	0.004	0.007	0.000	0.000	0.001
<i>GALNT3</i>	rs16851009	0.000	0.040	0.060	0.000	0.004	0.006	0.000	0.000	0.001
<i>PRDM2</i>	rs2245218	0.000	0.046	0.051	0.000	0.005	0.005	0.000	0.000	0.001
<i>PASD1</i>	rs7878232	0.001	0.040	0.027	0.000	0.004	0.003	0.000	0.000	0.000
---	rs1509269	0.000	0.024	0.029	0.000	0.003	0.003	0.000	0.000	0.000
---	rs11737074	0.000	0.019	0.016	0.000	0.002	0.002	0.000	0.000	0.000

The value 0.000 corresponds to estimated credibility &lt;0.001

# Correcting for possible bias

Let us consider that bias can cause an  $x$  proportion of variants pass a given p-value threshold for a specific phenotype association. If  $k$  variants have been tested, then the expected number of variants that pass the threshold due to bias is  $xk$ . If  $n$  variants have passed this threshold, then  $xk$  out of  $n$  are expected to reflect bias. By default, we don't know which these specific “biased” variants are. However, we can correct the credibility of each of the  $n$  variants for bias on average, multiplying by  $(n-xk)/n$ . For variants with uncorrected credibility estimates exceeding 50%, the corrected for bias credibility will remain above 50% if  $xk < (C-0.5)n/C$ .



# A few more questions

- “Conglomerate” evidence – e.g. various combinations of scattered studies, retrospective meta-analyses, prospective consortia analyses, in various time sequence: consider the highest level of evidence (what if the the lower levels are the large majority?)
- Tail of small studies with inflated effects causing heterogeneity: the evidence may grade “weak” overall, but the evidence of large studies may be graded “strong” overall
- Different genetic models (allele-based, recessive, dominant, etc): they may well be graded and they may get different grades in each of the three criteria, perhaps even in the overall grade; this is OK and even interesting to study
- Carefully defined subsets (racial descent, design or type of evidence based) may also be graded